# Open Challenges of Interactive Video Search and Evaluation

Jakub Lokoč
jakub.lokoc@matfyz.cuni.cz
Charles University
Prague, Czech Republic

Klaus Schoeffmann
ks@itec.aau.at
Alpen-Adria-Universität Klagenfurt
Klagenfurt, Austria

Werner Bailer
werner.bailer@joanneum.at
JOANNEUM RESEARCH
Graz, Austria

Luca Rossetto
rossetto@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Björn Þór Jónsson
bjorn@ru.is
Reykjavik University
Reykjavík, Iceland
IT University of Copenhagen
Copenhagen, Denmark

## ABSTRACT

During the last 10 years of Video Browser Showdown (VBS), there were many different approaches tested for known-item search and ad-hoc search tasks. Undoubtedly, teams incorporating state-of-the-art models from the machine learning domain had an advantage over teams focusing just on interactive interfaces. On the other hand, VBS results indicate that effective means of interaction with a search system is still necessary to accomplish challenging search tasks. In this tutorial, we summarize successful deep models tested at the Video Browser Showdown as well as interfaces designed on top of corresponding distance/similarity spaces. Our broad experience with competition organization and evaluation will be presented as well, focusing on promising findings and also challenging problems from the most recent iterations of the Video Browser Showdown.

## KEYWORDS

Video Retrieval, Interactive Retrieval, Retrieval Evaluation

**ACM Reference Format:**
Jakub Lokoč, Klaus Schoeffmann, Werner Bailer, Luca Rossetto, and Björn Þór Jónsson. 2022. Open Challenges of Interactive Video Search and Evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3503161.3546973

## 1 INTRODUCTION

The multimedia analysis and retrieval area has been significantly affected by the rise of deep learning [3]. Thousands of papers report on improvements of various deep architecture designs for particular multimedia data problems. However, all the reported results in benchmarks show just a limited picture of the real "out-of-the-box" performance. Even with strictly separated training and testing sets, good performance in a benchmark does not guarantee a success of the model for other data distributions. From this perspective, different types of comparative experiments are highly important to collect more evidence of the real performance of new and excitingly effective multimedia search models. Another issue emerges with a high number of similar items in a dataset, which affects complexity of specific search tasks as well. For example, what if users want to find one particular wedding party in a very large collection of wedding videos? Even a human-level classification performance for concepts like "wedding", "party", or "roses" does not guarantee effective search in such cases.

Moreover, users often have problems with selecting the proper class or object name, for various reasons. First, some models have a limited number of object/class categories; for example, the default YOLO v4 [1] model provides 80 COCO classes [10], but only a subset of these classes may be relevant for the search tasks and underlying dataset. Also, some classes have so many matches that they cannot be used reasonably (e.g., the *Person* object is easily found in tens of thousands of keyframes in the V3C1 dataset [12]). Contrarily, some other models provide so many different classes (e.g., when trained on ImageNet [2] with more than 21 000 classes) that users are overwhelmed by the number of classes and have difficulties in selecting the proper one. The latter issue is often also related to missing context or vocabulary of non-native speakers as users.

Hence, one of the most important types of evaluations in the multimedia search area is to incorporate user actions during the search process. Indeed, users may find it difficult to use a particular interface, or can overlook relevant data in current result sets, making automatic types of evaluations mere estimators of the real performance. On the other hand, if the current results do not lead to desired objects, users can positively affect the next steps of the search process. In order to inspect these search perspectives, several interactive multimedia search evaluation campaigns were established in the community. For example, the Video Browser Showdown competition (VBS) has celebrated its 10th anniversary in 2021 [6], while the Lifelog Search Challenge (LSC) was already organized for the fifth time [4]. Both competitions allow a comparison of approaches available at the event time, while winning systems (e.g., SOMHunter [9] in 2020, vitrivr [5] in 2021, or Vibro [7] in 2022) indicate potentially effective compositions of approaches to solving the competition tasks.

In the tutorial, we present and share our broad experience with participation at, and organization of, such interactive video search competitions. A wide spectrum of open challenges encountered so


*MM '22, October 10–14, 2022, Lisboa, Portugal*

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3546973

far will be presented, along with already established mechanisms for search task selection, presentation, and (remote) evaluation. Guidelines for designing a novel interactive video search system, as well as to improving an existing open-source system, will be provided in addition to a summary of promising search models identified so far at the competitions.

## 2 TUTORIAL CONTENTS

**Introduction.** We start our tutorial with a motivation of why interactive video search is still a hot topic, despite the high accuracy of deep learning models for automatic video content analysis in popular benchmark datasets. In particular, we show examples where automatic content analysis is not enough and where users need more than standard text-based querying. We discuss how to make fair and reproducible evaluations of different interactive video search (IVS) systems and give an overview of two popular evaluation campaigns, namely the Video Browser Showdown and the Lifelog Search Challenge.

**Tasks and Challenges.** We give an overview of different types of tasks that are evaluated at VBS and LSC, and what practical situations are modeled by them. We also talk about the simulated models behind these tasks and their limits, as well as which challenges may arise during evaluations. In this context, we discuss how to approach these challenges of participants and organizers and make sure that the evaluation results really reflect the true performance of search systems.

**Where is Deep Learning Helpful?** Every year, newly trained deep models emerge for various multimedia analysis tasks. With better multimedia content analysis, more effective ranking models and search systems can be designed for tasks evaluated at VBS/LSC. In this part of the tutorial, we summarize selected deep learning based models that enable effective meta-data extraction from contents, as well as extraction of useful content-based features for similarity modeling. Specifically, we show selected models for shot boundary detection, effective joint embedding for cross-modal similarity search, and approaches for additional content-based analysis and visualizations.

**Where Does Deep Learning Still Face Limitations?** Although new deep models integrated to VBS systems positively affected their performance, there are still limitations which must be taken into account. For example, users often find it difficult to describe some types of objects or events with a set of supported keywords. In other words, having a keyword query interface is not enough to solve a search task for users not familiar with the search task domain and without an experience with the classification model. Furthermore, even with proper keywords or a free-form text search option, users can still face an uneven distribution of concepts in the dataset. Either the candidate result set is too large, or the utilized words were not sufficiently represented in the train dataset of the employed model. We show examples of these limitations and shortcomings when using a respected search system for the V3C1 dataset.

**Evaluating Implementation Choices.** Understanding effects of various system design choices is particularly difficult for interactive systems. User studies, even for gathering interaction traces, are not amenable to such performance analysis, due to their excessive cost and ad-hoc nature. We present a methodology for using competition tasks to derive simulated users, whose interaction strategies can be varied systematically to expose their impact on system performance tradeoffs. We present recent application of these ideas for an interactive learning approach, and show how the methodology can be used to demonstrate scalability of such a system [8]. The concepts, however, should be applicable for a variety of interactive approaches.

**Comparative Evaluation Approaches and Mechanisms.** Interactive retrieval poses unique challenges not generally observed in most benchmarking campaigns. Specifically, system performance cannot be evaluated in isolation but rather requires end-to-end human-in-the-loop settings, where several systems and their users are brought together and evaluated in a controlled and comparable environment. Evaluation campaigns such as VBS or LSC achieved this over several years by physically bringing all participants to the same room, where the environment in which retrieval tasks are to be solved could be sufficiently controlled. Difficulties in international travel in recent years however necessitated the development of new approaches to support the evaluation of interactive retrieval solutions in a distributed setting.

We provide a special focus on the challenges and opportunities that come with conducting such evaluations in a fully distributed setting. This will also showcase the 'Distributed Retrieval Evaluation Server' (DRES) [11], an open-source retrieval evaluation infrastructure that currently powers both VBS and LSC.

With human-in-the-loop systems, ensuring reproducible results of evaluations becomes a non-trivial task. It requires not only suitable benchmark datasets, but also elaborate logging methods to ensure that all questions of interest can be answered based on the evaluation. Based on the dataset currently in use for VBS as well as several TRECVID tasks [12], we discuss various challenges in dataset design such as appropriateness of size, technical and semantic diversity of content, or legal limitations for distribution and re-use. We also introduce ongoing efforts in metrological and instrumentation developments aimed at gaining more insights into the interactive search process.

**Fun Test Session.** A real evaluation competition with several different VBS systems will be performed in this last part of the tutorial. This will give participants a hands-on experience and showcase the current state-of-the-art in IVS.

## ACKNOWLEDGMENT

This work has been supported by Charles University grant SVV-260588 and by the European Union's Horizon 2020 research and innovation programme, under grant agreement n° 951911 AI4Media (https://ai4media.eu).

## REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR* abs/2004.10934 (2020). arXiv:2004.10934 https://arxiv.org/abs/2004.10934

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[4] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus

Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *ICMR '22: International Conference on Multimedia Retrieval, Newark, NJ, USA, June 27 - 30, 2022*, Vincent Oria, Maria Luisa Sapino, Shin'ichi Satoh, Brigitte Kerhervé, Wen-Huang Cheng, Ichiro Ide, and Vivek K. Singh (Eds.). ACM, 685–687. https://doi.org/10.1145/3512527.3531439

[5] Silvan Heller, Ralph Gasser, Cristina Illi, Maurizio Pasquinelli, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2021. Towards Explainable Interactive Multi-modal Video Retrieval with Vitrivr. In *MultiMedia Modeling*. Springer International Publishing, Cham, 435–440.

[6] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoc, Andreas Leibetseder, Frantisek Mejzlík, Ladislav Peska, Luca Rossetto, Konstantin Schall, Klaus Schoeffmann, Heiko Schuldt, Florian Spiess, Ly-Duyen Tran, Lucia Vadicamo, Patrik Veselý, Stefanos Vrochidis, and Jiaxin Wu. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *Int. J. Multim. Inf. Retr.* 11, 1 (2022), 1–18. https://doi.org/10.1007/s13735-021-00225-2

[7] Nico Hezel, Konstantin Schall, Klaus Jung, and Kai Uwe Barthel. 2022. Efficient Search and Browsing of Large-Scale Video Collections with Vibro. In *MultiMedia Modeling*. Springer International Publishing, Cham, 487–492.

[8] Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2021. Impact of Interaction Strategies on User Relevance Feedback. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*. ACM, 590–598.

[9] Miroslav Kratochvíl, Frantisek Mejzlík, Patrik Veselý, Tomás Soucek, and Jakub Lokoc. 2020. SOMHunter: Lightweight Video Search System with SOM-Guided Relevance Feedback. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4481–4484. https://doi.org/10.1145/3394171.3414542

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[11] Luca Rossetto, Ralph Gasser, Loris Sauter, Abraham Bernstein, and Heiko Schuldt. 2021. A System for Interactive Multimedia Retrieval Evaluations. In *MultiMedia Modeling*, Jakub Lokoč, Tomáš Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras (Eds.). Springer International Publishing, Cham, 385–390.

[12] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. 2019. V3C–A Research Video Collection. In *International Conference on Multimedia Modeling*. Springer, 349–360.